DOCUMENT RESUME

ED 401 597 EA 027 901

AUTHOR Greene, Jay P.; And Others
TITLE The Effectiveness of School Choice in Milwaukee: A Secondary Analysis of Data from the Program's Evaluation. Occasional Paper 96-3.
PUB DATE Aug 96
NOTE 48p.; Paper presented at the Annual Meeting of the American Political Science Association (San Francisco, CA, August 30, 1996).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Academic Achievement; Data Analysis; Elementary Secondary Education; *Evaluation Problems; Program Effectiveness; Program Evaluation; Research Methodology; Research Problems; *School Choice; Scores; Standardized Tests; Statistical Analysis
IDENTIFIERS Milwaukee Public Schools WI

ABSTRACT

In 1990, Milwaukee (Wisconsin) became the site of the first publicly funded school-choice program providing low-income parents with vouchers that could be used to send their children to secular, private schools. An evaluation of Milwaukee's school-choice experiment was conducted by a team of researchers, headed by John Witte at the University of Wisconsin at Madison, during the years 1991-95. That study concluded that choice was not an effective way to improve the education of low-income, central-city students. The data were made available on the World Wide Web in February 1996. This paper presents findings of a study conducted by the Center for Public Policy at the University of Houston (CPP) and the Program in Education Policy and Governance at Harvard University (PEPG) that analyzed the University of Wisconsin-Madison database and research methodology. The CPP/PEPG study examined student performance as measured by standardized mathematics and reading tests. It concludes that students enrolled in choice schools for 3 or more years, on average, did better on standardized tests than a comparable group of students attending Milwaukee public schools. The results indicate that the reading scores of choice students in their 3rd and 4th years were, on average, from 3 and 5 percentile points higher, respectively, than those of comparable public school students. Math scores, on average, were 5 and 12 percentile points higher for the 3rd and 4th years, respectively. The CPP/PEPG study also argues that the earlier researchers failed to use analytic techniques appropriate to experimental data; the bulk of their research focused on comparisons between choice students and a much less disadvantaged cross-section of public school students. Nine tables are included. (Contains 30 end notes.) (LMI)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 401 597

EA

The Effectiveness of School Choice in Milwaukee: A Secondary Analysis of Data from the Program's Evaluation

Jay P. Greene
University of Houston

and

Paul E. Peterson
Harvard University

and

Jiangtao Du
Harvard University

with

Leesa Boeger and Curtis L. Frazier
University of Houston

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Program in Education Policy and Governance

HARVARD UNIVERSITY

Occasional Paper 96-3/August 1996

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
P. Peterson
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

EA027901

BEST COPY AVAILABLE

The Effectiveness of School Choice in Milwaukee:

A Secondary Analysis of Data from the Program's Evaluation

by

Jay P. Greene

Center for Public Policy

University of Houston

and

Paul E. Peterson

Program on Education Policy and Governance

John F. Kennedy School of Government and Department of Government

Harvard University

and

Jiangtao Du

Department of Statistics

Harvard University

with

Leesa Boeger and Curtis L. Frazier

Center for Public Policy

University of Houston

August 14, 1996

Paper prepared for presentation before the Panel on the Political Analysis of Urban School Systems at the August-September 1996 meetings of the American Political Science Association, San Francisco, California, Friday, August 30. Copyright reserved.

The Effectiveness of School Choice in Milwaukee:

A Secondary Analysis of Data from the Program's Evaluation

Executive Summary

In 1990 Milwaukee became the site of the first publicly funded school choice program providing low-income parents with vouchers that could be used to send their children to secular, private schools.* Milwaukee's school choice experiment was evaluated by a research team headed by political scientist John Witte at the University of Wisconsin at Madison. In five annual reports issued between 1991 and 1995, the researchers (hereinafter referred to simply as Witte) reported on the effectiveness of the Milwaukee experiment, as measured by the performance of students on standardized mathematics and reading tests. The senior author has summarized the results of his investigation as follows: "This school experiment . . . [has] not yet led to more effective schools. . . . Choice creates enormous enthusiasm among parents . . . but student achievement fails to rise."

Since this evaluation, until now, provided the only source of information on the test performance of choice students, many scholars, groups and foundations, drawing upon its findings, have concluded that school choice is not an effective way of improving

---

* Valuable technical advice was provided by Christopher Jencks, Robert Erikson, Frederick Mosteller, Donald Rubin, Kent Tedin, and Gregory Weiher. We are especially grateful to Donald Rubin for his detailed advice with respect to the analysis of data from a randomized experiment. Research assistance was provided by Chad Noyes and Jennifer Hill. The authors alone are responsible for the findings and conclusions reported herein.

the education of low-income, central-city students. The Carnegie Foundation for the Advancement of Teaching declared: "Milwaukee's plan has failed to demonstrate that vouchers... can spark school improvement." Albert Shanker, president of the American Federation of Teachers, claimed that the "private schools [in the Milwaukee choice plan] are not outperforming public schools."

For five years the researchers did not release data from the evaluation for secondary analysis by other members of the scholarly community. But in February of 1996 they made the data available on the World Wide Web. Over the past several months the Center for Public Policy at the University of Houston (CPP) and the Program in Education Policy and Governance at Harvard University (PEPG) have accessed the data, cleaned them of identifiable errors, and organized them into a readable usable format.

Although the certainty with which conclusions may be drawn is restricted by certain data limitations, results based upon the highest quality information contained within the data set indicate that attendance at a choice school for three or more years enhances academic performance, as measured by standardized math and reading test scores. Correcting for errors in the data set and using appropriate analytical techniques, the CPP/PEPG analysis of student performance finds that students enrolled in choice schools for three or more years, on average, do better on standardized tests, than a comparable group of students attending Milwaukee public schools.

The results indicate that the reading scores of choice students in their third and fourth years, were, on average, from 3 and 5 percentile points higher, respectively, than those of comparable public school students. Math scores, on average, were 5 and 12 percentile points higher for the third and fourth years, respectively. These differences are substantively significant. If similar success could be achieved for all minority students nationwide, it could close the gap separating white and minority test scores by somewhere between one-third and more than one-half.

CPP/PEPG results are based on data derived from a natural experiment that randomly assigned students to a test and control group. The natural experiment was the product of a mandate imposed on the program by the Wisconsin state legislature. It required choice schools, if oversubscribed, to admit applicants at random. This mandate created two randomly selected groups of students, one selected to participate in the choice program, the other not selected. The experimental situation is not unlike that widely practiced in medical research, where individuals are randomly allocated to treatment and control groups. The data are thus quite well suited for drawing scientific conclusions about the effectiveness of the choice program, provided they are analyzed correctly and interpreted cautiously.

The earlier analysis of the Milwaukee choice program did not give careful attention to this experimental data. On the one occasion when the experimental data were examined, the

researchers failed to employ appropriate analytical techniques. The bulk of their research efforts focused instead on comparisons between choice students and a much less disadvantaged cross-section of public school students. No valid conclusions can be drawn from the comparisons they conducted.

The Effectiveness of School Choice in Milwaukee:

A Secondary Analysis of Data from the Program's Evaluation

Milwaukee has for several years become the focus of attention for those concerned about education reform.* In 1990 Milwaukee became the site of the first publicly funded school choice program providing low-income parents with vouchers that could be used to send their children to secular, private schools. In 1995 the Wisconsin state legislature voted to expand the program to include religious schools, but the expanded program has been enjoined while constitutional issues are being resolved in the Wisconsin courts. Until then, the 1990 program, though limited in scope, remains the one opportunity to determine whether a government-sponsored program of school choice involving private schools can improve the educational performance of low-income, inner-city, minority children.[1]

Milwaukee's school choice experiment was evaluated by a research team headed by political scientist John Witte at the University of Wisconsin at Madison. In five annual reports issued between 1991 and 1995, the researchers (hereinafter referred to as Witte) reported on the effectiveness of the Milwaukee experiment, as measured by the performance of students on standardized mathematics and reading tests.[2] In each report

---

* Valuable technical advice was provided by Christopher Jencks, Robert Erikson, Frederick Mosteller, Donald Rubin, Kent Tedin, and Gregory Weiher. We are especially grateful to Donald Rubin for his detailed advice with respect to the analysis of data from a randomized experiment. Research assistance was provided by Chad Noyes and Jennifer Hill. The authors alone are responsible for the findings and conclusions reported herein.

the researchers concluded that students attending choice schools did not perform any better than did comparable students attending public schools. In a recently published article the senior author summarized the results of his investigation as follows: "This school experiment . . . [has] not yet led to more effective schools. . . .Choice creates enormous enthusiasm among parents . . . but student achievement fails to rise."[3]

Since this evaluation, until now, provided the only source of information on the test performance of choice students, many scholars, groups and foundations, drawing upon its findings, concluded that school choice is not an effective way of improving the education of low-income, central-city students. The Carnegie Foundation for the Advancement of Teaching has declared, "Milwaukee's plan has failed to demonstrate that vouchers... can spark school improvement."[4] Albert Shanker, president of the American Federation of Teachers claimed that the "private schools [in the Milwaukee choice plan] are not outperforming public schools."[5] The Texas State Teachers Association, a National Education Association affiliate, has avowed that "the results [in Milwaukee] have been dismal -- test scores have actually declined."[6] Harvard School of Education Professor Richard Elmore asserted that "thousands of children have participated in Milwaukee's public-private voucher experiment..., yet we see no discernible gains in learning."[7] The head of Wisconsin's leading teacher organization echoed these sentiments: "The bottom line ought to be whether kids learn more... and if you gauge it

by that, it doesn't measure up."[8] All these assessments depended upon the Witte study.

For five years the data from this evaluation were unavailable for secondary analysis by other members of the scholarly community. But in February of 1996 the data were made available on the World Wide Web. Over the past several months the Center for Public Policy at the University of Houston (CPP) and the Harvard Program on Education Policy and Governance (PEPG) accessed the data, cleaned them of identifiable errors, and organized them into a readily usable format.[9]

After correcting for detectable errors and using appropriate analytical techniques, CPP/PEPG found that students enrolled in choice schools for three or more years substantially outperformed, on average, a comparable group of students attending Milwaukee public schools. Although the certainty with which the conclusions may be drawn is restricted by certain data limitations, the CPP/PEPG analysis, using techniques appropriate to the analysis of experimental data, indicates that attendance at a choice school enhances academic performance, as measured by standardized test scores.

The Most Informative Data in the Evaluation

The bulk of the information on test scores that the earlier researchers collected is of marginal scientific value, because it only allows comparisons among decidedly different groups of students, a topic discussed later in this paper. But contained within the evaluation are data derived from a natural experiment

that assigned students at random to test and control groups. The data are thus quite well suited for drawing scientific conclusions about the effectiveness of the choice program, provided they are analyzed correctly and interpreted cautiously.

When enacting the legislation that gave rise to the choice program, the Wisconsin state legislature established conditions that allowed for a natural randomized experiment. The legislature required choice schools, if oversubscribed, to admit applicants at random. The requirement created two randomly selected groups of students, one selected to participate in the choice program, the other not selected. This experimental situation is not unlike that widely practiced in medical research where individuals are randomly allocated to treatment and control groups. Since the allocation is done at random, the two groups can be assumed to be similar, on average, in all respects other than the treatment. Any outcome differences can be reasonably attributed to the experimental condition.

In the field of education, random assignment rarely occurs, in part because it is difficult to justify denial of an educational benefit to children simply for purposes of educational experimentation. In Witte's original proposal to undertake the evaluation of the Milwaukee choice plan, he emphasized the unique research opportunity created by the legislative mandate requiring random acceptance:

> The students who applied but were not admitted will constitute the second group we will study. This

> is a unique opportunity in that it allows us to track students who will remain in the public schools but who come from families who have made an effort to seek private education. . . . By tracking the parallel educational outcomes of admitted and rejected students, we will have considerably improved control of a families' value of education.[10]

In one of his annual reports, Witte repeats this argument:

> Students not selected into the Choice Program in the random selection process represent a unique research opportunity. . . . If there are any unmeasured characteristics of families seeking private education, they should on average be similar between those in and not in the program.[11]

To exploit this research opportunity, the researchers collected data on the test performances and family background characteristics of students randomly selected into the choice program as well as those not selected.

When properly analyzed, these data indicate that choice students, when they remain in the choice experiment for three to four years, learn more than those not selected. The results indicate that the reading scores of choice students in years three and four, were, on average 3 and 5 percentile points higher, respectively, than those of the control group. Math scores were, on average 5 and 12 percentile points higher, respectively. These gains are not trivial. If similar success

could be achieved for all minority students nationwide, it could close the gap separating white and minority test scores by somewhere between one-third and more than one-half.[12]

Comparing Randomly Selected and Non-selected Students

These findings emerge from the CPP/PEPG analysis that takes into account the particular way in which the legislatively-mandated random assignment policy was implemented. Students did not apply to the choice program as a whole; instead, they applied each year for a seat in a particular grade in a particular school. They were selected or not selected randomly by school and by grade. Because the random assignment policy was implemented in this way, the analysis takes into account the year each student applied, the grade to which each student applied, and, to some extent, the particular school to which each student applied.[13]

The evaluation data distributed on the World Wide Web allowed the CPP/PEPG research team to take into account the grade to which the student applied and the year of application.[14] Nonetheless, CPP/PEPG was to some extent able to control for specific school effects by taking into account the ethnicity of the applicant. Over 80 percent of the choice students attended one of three schools, and, of these three schools, virtually all students applying to one school were Hispanic, while virtually all students applying to the two others were African American. As a result, we were, at least to some extent, able to estimate

the school to which a student applied by knowing whether they were Hispanic or African American. (Since the number of white students and other minority students for which information was available was so sparse that no reliable results could be obtained, these students were deleted from the analysis.)

Admission to the program was assumed to be at random for each of two ethnic groups, Hispanic and African American, for each of nine grades (K through 8) for each of four application years (1990 to 1993). This created 2 by 9 by 4 or 72 potential categories or points of comparison between those randomly selected into choice and those not selected (See Table 1).[15] The actual number of categories or blocks in any given analysis depends upon there being at least one observation within a block. By using standard statistical techniques for analyzing randomized block experimental data, with analysis of covariance or, similarly, regression adjustment to control for background characteristics, it was possible to estimate the effects of enrollment in choice schools on test scores. The procedure treats each block as a dummy variable in a regression equation that also includes the treatment variable and background characteristics.

The measures of test score performance are the same as the ones used in previous analyses, except for corrections of obvious errors. They consist of the students' national curve equivalent (NCE) scores for math and reading on the Iowa Test of Basic Skills. NCEs are derived from the national percentile rankings,

which place a student's performance relative to others taking the test, with a score of 60 indicating the student did better than 60 percent of others taking the test. The NCE is a transformation of the national percentile rankings that arranges the scores around the fiftieth percentile in a manner that can be described by a normal curve. A standard deviation for NCEs near the mean score is approximately 16 percentile points.

The comparison of selected and non-selected students was hampered by the fact that test data were available for only 76.2 percent of the selected students and 58.7 percent of the non-selected students (see Table 2). The analysis depends on the assumption that the missing cases do not differ appreciably from those remaining in the sample. One way of estimating whether this assumption is reasonable is to see whether the characteristics of selected students and non-selected students are similar.

The ethnicity, gender and initial test scores of the two groups do not differ in important respects. Neither do the two groups seem to differ on most other characteristics, if one assumes that parents who filled out questionnaires are representative of the overall population from which they were selected (see table 3).[16] Witte agrees: "In terms of demographic characteristics, non-selected . . . students came from very similar homes as choice [students did]. They were also similar in terms of prior achievement scores and parental involvement."[17]

The Results from Analysis of Experimental Data

Using the analytical procedures discussed above, CPP/PEPG estimated the effect of choice schools on student performance after one, two, three and four years in a choice school.[18] Using the blocking technique, the choice student test scores, after controlling for gender, are essentially compared with those of non-selected students who had applied the same year for the same grade and were of the same ethnicity.

The results of the main analysis are contained in table 4. They indicate that the effects of choice schools on test performance were trivial for the first two years students were in the program. But in year three and four choice students made substantial gains. On the math test, choice students scored, on average, 5 percentile points higher than non-selected students in year three and over 11 points higher in year four. On the reading test, choice students scored, on average, 3 percentile points higher after three years than those not selected into the program. After 4 years they scored nearly 5 percentile points higher. Statistical tests suggest that one can be confident that positive results of this magnitude would not appear, had choice schools had no effect.[19]

Controlling for Family Background

Data collection problems limit the extent to which the analysis can take into account family background characteristics. This poses no difficulty as long as it may be assumed that individuals in the analysis have been allocated at random to the

test and control groups. But given the fact that an appreciable number of cases are missing, it is possible that the two groups are no longer similar in all respects, despite their similar demographics. To see whether the results remain the same when background characteristics are taken into account a second analysis was performed (see table 5).

This analysis depends upon the information on family background characteristics obtained from parents at the time of the student's application to the choice program. Both test and questionnaire data were available from 36.7 percent of the families of selected students and 21.8 percent of the families of non-selected students (see table 2). Further reducing sample size was the fact that many parents did not respond to all the items in the questionnaire. As a result, the more family background characteristics that are controlled, the smaller the sample size. Controlling for additional background characteristics increases the precision of the analysis and adjusts for biasing differences. But these potential gains had to be weighed against the cost of losing still additional subjects from an analysis already diminished in size.

Balancing these considerations against one another, CPP/PEPG controlled for family income and mother's education but not for other family background characteristics. Past scholarly research has shown that family income and mother's education strongly affect a child's educational performance, and most parents who returned questionnaires responded to these two questions. The

response rate for other family background characteristics was considerably less, and inclusion of these additional variables would have further reduced the size of an already small sample.

When controls for family income and mothers' education are added to the variables included in the main analysis reported above, the sample size drops but the substantive findings change hardly at all (see table 5), a result to be expected if assignment to the treatment and control groups was truly at random, and non-response was similarly at random. The effects of choice on student performance after two years in a choice school were trivial and inconsistent. But after three and four years in a choice school, students scored noticeably higher than the non-selected students remaining in the public schools. On the math test, choice students scored an estimated 7 percentile points higher in year three and an estimated 10 points higher in year four. On the reading test, choice students in their third year outperformed the control group by an average of 6 points on the reading test; in the fourth year they scored an estimated 4 percentile points higher. Because this analysis depends upon a sample much smaller in size, the results do not achieve the same level of statistical significance as do the results for the main analysis. Yet the consistency of the estimated effects generated by the two analyses lends weight to the conclusion that enrolling in choice schools yields decidedly positive effects after a students' third and fourth years.

Controlling for Prior Test Scores

The results reported above do not control for student test scores prior to entry into the choice program. It is not necessary to control for prior test scores when comparing a test group and a control group in an experimental situation, because the two groups, if randomly assigned to each category, can be assumed to be similar. But because of the sizeable number of missing cases, it is possible that the selected and non-selected groups included in the analysis differed in this important respect.

This potential source of bias did not appear, however. The average test scores at the time of application for the two groups was essentially the same. The average math and reading test scores for those selected into choice were the NCE equivalent of a 39 and 38 percentile ranking, respectively; for those not selected they were at the 39 percentile for reading and 40th for math (see table 3).

Since the test scores at the time of application were essentially the same, it was unlikely that controls for this variable would alter the result. CPP/PEPG nonetheless tested for the possibility and the results are reported in Table 6. Because test scores at the time of application were available only for a limited number of applicants, the sample size for this test was reduced. Yet with only one exception--the fourth year reading results based on just 26 observations--the results controlling

for prior test scores do not differ substantially from the results reported in the main analysis reported in Table 4.

Checking for Selection Effects

The main analysis indicates that choice has sizeable, statistically significant effects in the third and fourth years of a student's program. The observed effects may be hypothesized as being produced by two quite different processes:

1) Students benefit in measurable ways from the choice experience only after participating in the program for three or more years.

2) Students remain in the program for three to four years only if they have benefitted from the experience.

To ascertain whether the effects observed in the main analysis were due to processes suggested by the first or second hypothesis, CPP/PEPG analyzed the effects of choice on the first and second year scores of only those students for whom test results are available in years three and four. The effects of choice on their performances during the first two years differ but little from the effects for all first and second year students (see table 7). These results suggest that the substantial effects of choice schools in years three and four are, on the whole, not due to differential student retention

rates but to accumulated learning over the three-to-four year period of time.

Interpreting the Findings

During their first two years in a choice school, students in choice schools performed similarly on math and reading tests comparable students attending Milwaukee public schools. But in their third and fourth years the performances of students in choice schools were noticeably superior to those of similarly-situated students in Milwaukee public schools. The results are quite consistent with a common-sense understanding of the educational process. Choice schools are not magic bullets that transform children overnight. It takes time to adjust to a new teaching and learning environment. The disruption of switching schools and adjusting to new routines and expectations may hinder improvement in test scores in the first year or two of being in a choice school. Educational benefits accumulate and multiply with the passage of time. One can hardly be surprised that their impact is felt only with the passage of time.

Why the Earlier Analysis Produced Different Results

In its fourth year report the earlier researchers reports their own analysis of the performances of selected and non-selected students.[20] The researchers find no significant choice-school effects on student performance. These findings depend on analytical techniques that fall well short of

appropriate statistical procedure for analyzing data from a randomized experiment.

First, the researchers failed to categorize or block their data so as to take account of the fact that random assignment was by grade, year of application, and by school. Instead, they simply categorized all the selected students together, then compared them to the entire group of non-selected students. They ignored the fact that the natural experiment was random only by grade, year and school. They failed to create a statistical model that approximated the actual character of the natural experiment. These analytical deficiencies contaminate their results.

Second, the researchers, in attempting to control for prior test scores, did not distinguish between test scores achieved before entry into choice schools from scores achieved after entry. In so doing, they attempted to estimate the effects of school choice while controlling for a portion of these effects. Also, instead of examining the effects of choice schools over the entire time period that students were exposed to the experimental condition, they measured the changes in test scores from one year to the next.

The egregious nature of these errors can be appreciated by imagining an experiment to determine if fertilizer helps corn mature faster.[21] In this experiment the farmer fertilizes one field before planting and monthly thereafter; in the other field the farmer fertilizes not at all. To measure the effect of the

fertilizer, the experimenter decides to calculates only how fast the corn grows from knee high to shoulder height, ignoring the possibility that much of the fertilizer's effect may have occurred prior to its reaching knee height. Similarly, controlling for test scores received during the first years in a choice school incorrectly controls for some of the effect one is seeking to test. It also incorrectly assumes that the rate of academic progress for students is, on average, a steady upward line not subject to irregular spurts.

How about all the other Data in the Evaluation?

Instead of using the best analytical techniques available for the analysis of experimental data, the Witte team report results from the analysis of non-experimental information that in their analyses tell us little, if anything, about the effectiveness of school choice. The study's analytical errors can be compared to the classic errors committed by the Literary Digest poll taken in 1936. The magazine tried to predict the outcome of the presidential election by mailing out a questionnaire to ten million Americans. Unfortunately, the 2.2 million people who responded were a group not representative of the American public. As a result, the Literary Digest, not realizing the biases in their sampling technique, predicted Alf Landon would win by 57 percent of the vote. When Roosevelt won by 62 percent of the vote, the Digest soon went out of business. Meanwhile, George Gallup, employing a scientific data collection

technique, accurately predicted a Roosevelt victory with information from fewer than two thousand citizens. It is now elementary to observe "a large sample is no guarantee of accuracy."[22] The results from the earlier study of school choice in Milwaukee, no matter how large its sample, have no more scientific validity than the poll conducted by the Literary Digest, because the population constituting their control group is not representative of choice students, had they remained in the public schools.

The earlier finding that the choice program has "not yet led to more effective schools" relies upon four flawed comparisons of choice-school students with Milwaukee public-school students (in addition to an incorrect analysis of the experimental data discussed above):

I. Comparison of test scores of cohorts of choice students with those of public school students.

II. Comparison of the changes from year to year in the individual test scores of choice students with public school students.

III. Comparison of the test scores of choice students with low-income students in the Milwaukee public schools.

IV. Analysis of the effect of choice by means of a multiple regression analysis of all changes in test scores.

All four techniques, as used by in the earlier study, are seriously flawed.

Method I: Cohort comparison

When comparing the test scores of yearly cohorts of choice students with cohorts of a sample of Milwaukee public school students, the researchers find no stable, significant difference between them. But even if they had, the finding would have been meaningless. Cohort comparison is perfectly appropriate if one is making comparisons between individuals randomly assigned to treatment and control groups. But one can hardly make such an assumption when comparing choice students to a sample of Milwaukee public school students, especially when the treatment and control group are different in many important respects (see table 9). Before entering the program, the soon-to-become choice students scored well below the average of a cross-section of public-school students. The average score on the math test of the choice student at the time of admission was at the 39th percentile, while the average public school student initially scored at the 45th percentile. The average reading score for the choice student was at the 38th percentile, while the average public school student's initial score was at the 43th. Since students had decidedly lower scores before entering choice schools, it is misleading simply to compare post-test scores. Witte, immediately after reporting the results, says this is not a "way to accurately measure achievement gains and losses."[23]

Method II: Comparing Changes in Test Scores

The researchers next compared annual changes in achievement scores of choice and public school students. In two of the six comparisons reported in their first published paper, the choice schools score higher; in the other four public schools do better.[24]

This kind of comparison incorrectly assumes that choice and public school students were similar in all respects other than their initial scores, an assumption that can appropriately be made in the case of a randomized experiment but which is in this case unwarranted. Choice students were not at all comparable to the public school students included in the comparison group. In fact, the choice students available for their analysis were different in many ways that may well be associated with lower test scores, including the following:

* Ninety-seven percent of choice students were African American or Hispanic, while only 60 percent of the public school control group were from these ethnic groups.
* Choice parents reported their family income to be $11,330 as compared to the $20,040 reported by the average Milwaukee public school parent.
* Only 24 percent of choice families reported being married; 47 percent of Milwaukee public school parents did.[25]

* Mothers of choice students were more likely to be receiving welfare assistance than were mothers of public-school students.

* On a scale where 4 indicates a high school graduate and 5 some college education, mothers of choice students report an average level of education of 4.2 compared to 3.9 for public-school mothers.

All but one of these differences, earlier research has shown, is likely to produce results in which choice students will appear to have achieved fewer test score gains than the comparison group.[26] The fact that choice students are significantly more likely to come from households headed by poor, minority, single mothers makes any comparison between them and Witte's public school sample highly misleading. As one of the earlier researchers admitted, "As for change scores, they are next to meaningless, since the bivariate comparisons don't control for any of the known differences between the groups."[27]

Method III: Comparing Scores of Low-Income Students

The previous researchers attempted to mitigate the problems associated with Methods I and II by also comparing cohort and change scores of choice students with public school students from low-income families. This approach suffers not only from not controlling the full array of family background characteristics but, quite specifically, from their use of a flawed measure of

family income. The measure of income used in the vast majority of the comparisons is whether or not a student receives free or reduced-cost school lunch.

The measure divides the population into two income groups: those who receive a subsidized lunch and those who do not. On its face, such a simple dichotomy is an entirely inadequate measure of family income. With only two categories, the variable inevitably lumps together people from unlike circumstances.

Even worse, the subsidized lunch measure seems to be an extremely inaccurate measure of family income. Several types of error may occur. Some families may not request a subsidized lunch, even though their household income would make them eligible. Other families may report low income in order to receive a government benefit, even though they are not eligible. It is also possible that claims are submitted on behalf of families by school officials anxious to ensure that all students receive their school lunch. Finally, the Milwaukee public school subsidized lunch records may be faulty.

Whatever the sources of error, it is not a trivial mistake. The subsidized lunch measure of family income has only a weak correlation with parental reports of household income, as reported in parental questionnaire. More than 16 percent of Milwaukee public-school families who report incomes over $42,500 are designated as receiving subsidized lunch. Meanwhile, 26 percent of choice students with family income below $17,500 did

not receive free lunch, compared to 14 percent of public school students.

It is also peculiar that the school lunch variable indicates that the incidence of low income is higher in public schools than among choice students. Only 69% of choice students are designated as receiving free lunch as compared to 74% of Milwaukee public school students. Yet average income of public-school parents reported was $20,040, while the average income reported by choice parents was only $11,330.

It is this flawed measure of income upon which the researchers depend for their many tables and regressions that compare choice and "low-income" Milwaukee public school students.

Method IV: The Regression Analysis

The fourth method used by the previous research team estimates the effects of choice after controlling for several family background characteristics, as reported by parents in a questionnaire.[28] This analysis reveals choice to have negative effects on reading scores and positive but insignificant effects on math scores.

Although this analysis no longer uses the subsidized lunch as its indicator of family income and attempts to take into account the many ways in which choice students differ from Milwaukee public school students, a number of serious problems remain, three worthy of special mention:

* First, the public school control group used in these regressions was in no way comparable to students in the choice program. Indeed, it was not even a random sample of the Milwaukee public school population. For this analysis, the researchers only included those students for whom both family background information and changes in test scores were available. Demographic information was obtained in a questionnaire distributed to a random sample of choice and public-school parents. Because of the very low response rate to the survey and spotty test-score records, less than 20 percent of the choice students were available for the regression analysis. Less than 10 percent of the public school students were included in the analysis (see table 8).

Those for whom the necessary information exists differ from non-respondents (see Table 9). Information supplied by the Milwaukee public schools shows that respondents were less likely to be of minority background and scored higher on both the math and reading tests. In short, this regression compares choice students to a self-selected group of public school students whose parents had responded to the questionnaire.

* Second, the researchers incorrectly "stacked" the data, a practice which combines all year-to-year changes into one analysis. Each student is counted as

an independent observation for each year in which they remain in the study. The units of analysis are not students, but student-years. But stacking data is inconsistent with a basic assumption of the regression analysis that each observation is independent.

"Stacking" data sets is also conceptually flawed. In the stacked regression the researchers attempt to predict each year's scores while controlling for the prior year's results. But the prior year's scores for choice students are already affected by participating in the choice program. To control for these scores the researchers controlled for program benefits while trying to measure them. Data stacking also makes the improbable assumption that students learn at uniform rates throughout the length of the program.

* Third, the regression analysis works with a data set that has a very large number of missing cases. It must make an estimate of choice-school effects with only 19 percent of his original number of cases and only 9 percent of his original number of public school student cases. By comparison the main CPP/PEPG analysis utilizes 76.2 percent of the original number of choice-student cases and 58.7 percent of the control group cases.

The test and control populations included in the regression analysis differ dramatically in almost every

respect--initial test score, race, income, household structure and educational attainment. In all but one respect, choice students are the disadvantaged group (see table 9). By comparison, the test and control groups in the CPP/PEPG analysis have similar demographic profiles (see table 3).

* Fourth, improbable assumptions must be made when using regressions comparing a test group to a much more heterogeneous control group (wider range of educational performance, greater ethnic diversity, wider range of incomes, etc.). Regression analysis must assume that relationships among these variables are identical over their entire range, and this is unlikely to be the case. For example, it must assume that the effects of a $5,000 increase in income are the same, regardless of whether the increase is from $10 to $15 thousand or $45 to $50 thousand.

Such an assumption is particularly problematic when test and control groups are extremely dissimilar, as is the case here. For example, the public school comparison group included many white students from families earning over $20,000 annually. There were virtually no equivalents in the choice sample. Under these circumstances, linear regression is being asked to perform an analysis for which the technique is poorly equipped.[29] When one divides the data set into

> contrasting income and racial groups, one finds that within each subgroup many variables included in the regression have different slopes and, on occasion, even different signs. When this occurs, it violates the an assumption upon which linear regression analysis depends. Any results based on such an analysis must be treated with extreme suspicion.

In short, the four attempts by earlier researchers to find out whether choice schools offer no evidence on the effectiveness of choice schools. As the professional staff of the Wisconsin Legislative Audit Bureau observed in its 1995 audit of the program,

> Professor Witte's conclusion, that there is no difference between the academic performance of students in choice schools and those in public school schools, stated in his fourth annual report in January 1995, is stronger than can be supported by the limited data available. In fact, no conclusion can be drawn . . . .[30]

Conclusions

The Milwaukee choice plan, approved by the Wisconsin state legislature in 1990, suffered from severe legislative restrictions that made it difficult for the program to succeed. Restrictions included the following:

* Only several hundred children from low-income households were eligible for choice.

* The voucher they received was worth half the cost of educating a student in the Milwaukee Public Schools.

* Parochial schools were excluded, preventing parents from choosing one of more than 90% of Milwaukee's private school options.

But despite these restrictions and limitations, data derived from a natural experiment that allocated students randomly to test and control groups suggests that students in choice schools, in their third and fourth years, scored, on average, from 3 to 5 percentile points higher in reading and 5 to 12 points higher in mathematics than a randomly selected control group. These are not trivial differences in educational achievement. A difference of eight points wipes out half the observed difference between the performance of whites and minorities on nationally standardized tests. If even this limited choice program has the capacity to make such an extraordinary contribution to equal educational opportunity, more extensive choice plans deserves far more serious consideration than they have generally received.

Because a significant number of cases are missing, one cannot draw conclusions with complete certainty. But despite data restrictions, an appropriate statistical analysis of data from a natural randomized experiment contradicts the findings of earlier research on the Milwaukee choice program. Instead of indicating that choice schools are not effective, as earlier scholars have claimed, the weight of the evidence points in exactly the opposite direction. The highest quality evidence in

the data set indicates that students in choice schools learn more after three to four years.

Table 1-- Blocks within which Selected and Non-Selected Students were Classified

	Year of Application			
Grade and Ethnicity	1990	1991	1992	1993
Kindergarten				
African American	1	2	3	4
Hispanic	5	6	7	8
Grade 1				
African American	9	10	11	12
Hispanic	13	14	15	16
Grade 2				
African American	17	18	19	20
Hispanic	21	22	23	24
Grade 3				
African American	25	26	27	28
Hispanic	29	30	31	32
Grade 4				
African American	33	34	35	36
Hispanic	37	38	39	40
Grade 5				
African American	41	42	43	44
Hispanic	45	46	47	48
Grade 6				
African American	49	50	51	52
Hispanic	53	54	55	56
Grade 7				
African American	57	58	59	60
Hispanic	61	62	63	64
Grade 8				
African American	65	66	67	68
Hispanic	69	70	71	72

Table 2-- Percentage of Parents of Selected and Non-selected Students With Test Scores and Responding to Parent Questionnaire

	Selected	Non-selected
Number of Applicants	1,356	693
Number of Students for which Test Score is Available	1,034	407
Percentage of Cases Included in CPP/PEPG Main Analysis	76.2	58.7
Number of Students for which Test Score and Parent Survey Data are Available	497	151
Percentage of Cases included in CPP/PEPG Second Analysis	36.7	21.8

Table 3--Differences Between Selected and Non-selected Students*

All Students for Which Tests Scores are Available	Selected Students	Non-Selected Students
Math Pre-test (Average)	39	40
Reading Pre-test (Average)	38	39
% Black	77	82
% Hispanic	20	13
% Male	44	52
Grade Applied	2.8	3.6
Students for which Both Test Score and Parent Survey Results are Available	**Selected Students**	**Non-selected Students**
Average Score on Prior Math Test	40	38
Average Score on Prior Reading Test	39	38
% Black	80	82
% Hispanic	17	15
% Male	45	51
% Married	24	32
% AFDC	57	55
Mother's Education (High School Diploma = 4)	4.2	3.8
Family Income	$11,250	$11,500
Grade Applied	2.7	3.5

* All data were blocked by ethnicity. Gender differences were controlled in the main analysis. Gender, education and income differences were controlled in the second analysis.

Table 4--**The Main Analysis:** Percentile Point Effect of Choice Schools on Student Performances on Standardized Tests, Controlling for Gender and Blocking Data by Ethnicity, Year of Entry and Grade Level

Effect of Choice School on Performance on . . .				
Iowa Tests of Basic Skills	**Years in Choice School**			
Mathematics Test	**First**	**Second**	**Third**	**Fourth**
Estimated Effect of Choice	-0.49	-0.87	4.98	11.59
Standard Error	(1.77)	(1.92)	(2.62)	(4.62)
P value < (1-tail test)	0.39	0.33	0.03	0.01
P value < (2-tail test)	0.78	0.65	0.06	0.01
Number of cases	727	568	310	110
Iowa Tests of Basic Skills	**Years in Choice School**			
Reading Test	**First**	**Second**	**Third**	**Fourth**
Estimated Effect of Choice	-0.13	-0.06	3.13	4.81
Standard Error	(1.55)	(1.68)	(2.21)	(4.17)
P value < (1-tail test)	0.47	0.49	0.08	0.13
P value < (2-tail test)	0.93	0.97	0.16	0.25
Number of cases	691	576	309	108

Table 5--Percentile Point Effect of Choice Schools on Student Performances on Standardized Tests, Blocking Data by Ethnicity, Year of Entry and Grade Level and Controlling for Gender, Family Income and Mother's Education

Effect of Choice School on Performance on . . .				
Iowa Tests of Basic Skills	**Years in Choice School**			
Mathematics Test	**First**	**Second**	**Third**	**Fourth**
Estimated Effect of Choice	3.59	1.16	7.07	9.90
Standard Error	(2.89)	(3.10)	(4.43)	(9.01)
P value < (1-tail test)	0.11	0.35	0.06	0.14
P value < (2-tail test)	0.22	0.71	0.11	0.28
Number of cases	361	291	161	63
Iowa Tests of Basic Skills	**Years in Choice School**			
Reading Test	**First**	**Second**	**Third**	**Fourth**
Estimated Effect of Choice	1.38	-3.06	5.80	4.04
Standard Error	(2.45)	(2.63)	(4.21)	(7.50)
P value < (1-tail test)	0.29	0.13	0.09	0.30
P value < (2-tail test)	0.58	0.25	0.17	0.59
Number of cases	338	297	160	60

Table 6-- Results for Only Those Students Tested Before Entering Choice Program: Percentile Point Effect of Choice Schools on Student Performances on Standardized Tests, Controlling for Gender and Test Prior to Entry and Blocking Data by Ethnicity, Year of Entry and Grade Level

Effect of Choice School on Performance on . . .				
Iowa Tests of Basic Skills	**Years in Choice School**			
Mathematics Test	**First**	**Second**	**Third**	**Fourth**
Estimated Effect of Choice	0.58	-0.61	9.07	10.14
Standard Error	(1.91)	(2.62)	(3.60)	(9.67)
P value < (1-tail test)	0.38	0.41	0.01	0.16
P value < (2-tail test)	0.76	0.82	0.01	0.31
Number of cases	319	171	86	26
Iowa Tests of Basic Skills	**Years in Choice School**			
Reading Test	**First**	**Second**	**Third**	**Fourth**
Estimated Effect of Choice	-0.94	-0.19	6.98	-0.39
Standard Error	(1.77)	(2.44)	(3.32)	(8.29)
P value < (1-tail test)	0.30	0.47	0.02	0.48
P value < (2-tail test)	0.59	0.94	0.04	0.96
Number of cases	327	174	87	26

Table 7-- Results for First Two Years of Students Remaining in Choice Compared to All Students: Percentile Point Effect of Choice Schools on Student Performances on Standardized Tests, Controlling for Gender and Blocking Data by Ethnicity, Year of Entry and Grade Level

Effect of Choice School on Performance on . . .

	Students Remaining in Choice		All Students (From Table 4)	
Iowa Tests of Basic Skills	**Years in Choice**		**Years in Choice**	
Mathematics Test	**First**	**Second**	**First**	**Second**
Estimated Effect of Choice	0.81	1.23	-0.49	-0.87
Standard Error	(3.00)	(2.46)	(1.77)	(1.92)
P value < (1-tail test)	0.39	0.31	0.39	0.33
P value < (2-tail test)	0.79	0.62	0.78	0.65
Number of cases	357	353	727	568

	Students Remaining in Choice		All Students (From Table 4)	
Iowa Tests of Basic Skills	**Years in Choice**		**Years in Choice**	
Mathematics Test	**First**	**Second**	**First**	**Second**
Estimated Effect of Choice	1.75	1.80	-0.13	-0.06
Standard Error	(2.64)	(2.20)	(1.55)	(1.68)
P value < (1-tail test)	0.26	0.21	0.47	0.49
P value < (2-tail test)	0.51	0.42	0.93	0.97
Number of cases	349	356	691	576

Table 8.--Percentage of Parents of Choice and Public School Students With Test Scores and Responding to Parent Questionnaire

	Choice Students	Public School Students
Original Sample	1,613	6,549
Number for whom Change in Test Scores can be Estimated	499	2,033
Number of Students for whom both Test Scores and Parental Survey Data is available	303	610
Percentage of Cases Included in Method IV Regression Analysis	19.0	9.3

Table 9: Differences Between Choice Students and Public School Students Included in Method IV Regression Analysis

Student Characteristics	Choice Students		Public School Students		
	All[a]	Analyzed[b]	All[a]	Analyzed[b]	p value <[c]
Average Score on Prior Math Test	39	40	45	49	.01
Average Score on Prior Reading Test	38	39	43	47	.01
% Black	74	81	59	50	.01
% Hispanic	21	16	11	10	.01
% Male	46	47	52	50	.16
% Married	---	24	---	47	.01
% AFDC	---	58	---	40	.01
Mother's Education (H.S. Diploma = 4)	---	4.2	---	3.9	.01
Family Income	---	$11,330	---	$20,040	.01

a All public school students in a randomly selected sample taken from the public school records. All choice students who were enrolled, except for test data, which was available for 71.8% of those enrolled.

b Students for whom parent questionnaire was filed and 2 test scores are available so that changes in test scores can be ascertained.

c Significance of difference between choice and public school students available for Method IV regression analysis.

ENDNOTES

1. Several studies suggest that privately-funded school choice programs enhance the performance of low-income, minority, inner-city students. See Terry M. Moe, ed., Private Vouchers (Stanford University: Hoover Institution Press, 1995.

2. John F. Witte, "First Year Report: Milwaukee Parental Choice Program." Department of Political Science and the Robert M. La Follette Institute of Public Affairs, University of Wisconsin-Madison. (November, 1991); John F. Witte, Andrea B. Bailey and Christopher A. Thorn. "Second Year Report: Milwaukee Parental Choice Program." Department of Political Science and the Robert M. La Follette Institute of Public Affairs, University of Wisconsin-Madison, Madison Wisconsin. (December, 1992.); John F. Witte, Andrea B. Bailey and Christopher A. Thorn, "Third Year Report: Milwaukee Parental Choice Program." Department of Political Science and the Robert M. La Follette Institute of Public Affairs, University of Wisconsin-Madison, Madison Wisconsin. (December, 1993.); John F. Witte, Christopher A. Thorn, Kim M. Pritchard, and Michele Claibourn, "Fourth Year Report: Milwaukee Parental Choice Program." Department of Political Science and the Robert M. La Follette Institute of Public Affairs, University of Wisconsin-Madison, Madison Wisconsin. (December, 1994). John F. Witte, Troy D. Sterr and Christopher A. Thorn, "Fifth-Year Report: Milwaukee Parental Choice Program," Department of Political Science and The Robert M. La Follette Institute of Public Affairs, University of Wisconsin-Madison, December 1995).

3. John Witte, "Who Benefits from the Milwaukee Choice Program?" in Richard Elmore, Bruce Fuller, and Gary Orfield, eds. Who Chooses? Who Loses?: Culture, Institutions and the Unequal Effects of School Choice (New York: Teachers College,Columbia University, 1996, pp. 118- 37.

4. Carnegie Foundation for the Advancement of Teaching, School Choice (Princeton, New Jersey: The Carnegie Foundation for the Advancement of Teaching), 1992, p. 73.

5. "Where We Stand: All Smiles," Advertisement in New York Times, January 30, 1994. Also see similar assessments by a spokesperson for the American Federation of Teachers quoted in Daniel McGroarty, "School Choice Slandered," The Public Interest, No. 117, (Fall, 1995), p. 96.

6. "Our Public Schools... The Best Choice for Texas: Issue: Public vs. Private," (Austin, Texas: Texas State Teachers Association, 1995).

7. Curtis Lawrence, "Choice Programs Showing Little Success, Study Says," Milwaukee Journal Sentinel, July 15, 1995, p. 7.

8. Laura Miller, "Wis. Vouchers for Religious School Urged," Education Week, January 25, 1995, p. 11.

9. This proved to be a more challenging task than originally anticipated. The data available on the World Wide Web are poorly organized and contain numerous errors. For example, we discovered that the test scores that provide the dependent variable in the Witte analysis, the National Curvilinear Equivalents (NCEs), were for choice students incorrectly computed from their national percentage ratings in as many as 80 percent of the cases in some years, despite the fact this computation can be made by a simple translation formula. Although these errors did not appear to be systematically biased, the introduction of even random error into the data set inflates standard errors and reduces the likelihood of finding significant effects.

10. John F. Witte, "Research Proposal: Milwaukee Parental Private School Choice Program," September 13, 1990, as quoted in George A. Mitchell, "The Milwaukee Parental Choice Program," (Milwaukee, Wisconsin: Wisconsin Policy Research Institute, Inc.), November, 1992, Vol 5, No. 5.

11. Witte, Thorn, Pritchard, and Claibourn, 1994, p. 24.

12. Many studies have found a difference of approximately one standard deviation between the average test scores nationwide of white and minority students. See, for example, Ina V. S. Mullis and Lynn B. Jenkins, The Reading Report Card, 1971 - 1988: Trends from the Nation's Report Card (Office of Educational Research and Improvement, U. S. Department of Education, January, 1990), pp. 56-65.

13. One of the authors of this paper earlier expressed the view that these characteristics of the admissions process made it impossible to conduct a valid comparison of the two groups. Paul E. Peterson, "A Critique of the Witte Evaluation of Milwaukee's School Choice Program." Center for American Political Studies Occasional Paper 95-2. Littauer Center, Department of Government, Harvard University, 1995. He is indebted to Donald Rubin for information showing the way in which one could still make valid comparisons by use of the "blocking" technique described in the text.

Siblings were exempt from the random assignment rule. We do not know the extent to which this exemption invalidates the findings reported in the text.

14. The stated grounds for the refusal was to protect the confidentiality of the students and the schools. CPP\PEPG offered to sign a form protecting the confidentiality of students, a standard practice within the education research community, but both Witte and the Wisconsin Department of Public Instruction remained unwilling to provide this information.

15. The analytical procedures are described and discussed in D. B. Rubin, "William G. Cochran's Contributions to the Design, Analysis, and Evaluation of Observational Studies," in Podurl S. R. S. Rao, W. G. Cochran's Impact on Statistics (New York: John Wiley & Sons, 1984: W. G. Cochran, "The Planning of Observational Studies of Human Populations," in J. Roy, Statistical Sociology 128 (12965) A: 234 - 65.

16. The use of confidence tests to see whether the two groups differ significantly in some statistical sense is in this case inappropriate, because comparisons need to be made for each block of data, within which sample sizes are too small for such a test to be appropriate. But the information in table 3 suggests that the two groups of students closely resemble one another.

17. Witte, Thorn, Pritchard, and Claibourn, 1994, p. 26.

18. These data are from the first four years of the choice school experiment. Public-school test score information was not available for subsequent years.

19. See table 4. We prefer the one-tailed t-test to estimate the statistical significance of the findings, because theory suggests that students should perform better in a choice school. Critics of choice typically prefer the null hypothesis that there are no effects of choice. But for the even more skeptical reader, who thinks one is as likely to obtain negative as positive effects, we also report a two-tailed t-test in table 4.

20. Witte, Thorn, Pritchard, and Claibourn, 1994, pp. 24 -26.

21. We are indebted to Donald Rubin for suggesting this analogy.

22. Robert S. Erickson, Norman R. Luttbeg and Kent L. Tedin, American Public Opinion (New York: Macmillan, 1991), p. 27.

23. Witte, 1996, p. 28.

24. Witte, 1996, pp. 128-30.

25. Witte, Thorn, Pritchard and Claibourn, 1994, Table 5C.

26. Christopher Jencks, Who Gets Ahead? The Determinants of Economic Success in America (New York: Basic Books, 1972).

27. Christopher A. Thorn, E-mail communication to Leesa Boeger, May, 7 1996. 09:47:20 Central Standard Time.

28. Witte, et al. "Fifth Annual Report," 1995, Tables 12a and 12b.

29. In the CPP\PEPG analyses regression analysis is possible because the test and control group are similar (see table 3).

30. Wisconsin Legislative Audit Bureau, An Evaluation of Milwaukee Parental Choice Program, State of Wisconsin, Legislative Audit Bureau, 1995, p. 31.